BREAST CANCER PREDICTION BY MACHINE LEARNING

Adarsh Reddy
Dept of Computer Science and Engineering
R.V.College of Engineering Benguluru-560059

Mallikarjun Kamble
Dept of Computer Science and Engineering
R.V.College of Engineering Benguluru-560059

Manu Bhardwaj
Dept of Computer Science and Engineering
R.V.College of Engineering Benguluru-560059

Prof. Sharvani GS
Assistant Professor
Dept of Computer Science and Engineering
R.V.College of Engineering Benguluru-560059

I. Introduction

Breast cancer disease is one of the most common cancers among women all around the world, representing most new cancer cases and cancer-related diseases are leading to deaths according to global statistics, making it a most significant public health problem in today’s society. The early diagnosis of Breast Cancer can improve the prognosis and chance of survival significantly, as it can suggest timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of Breast Cancer and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex Breast Cancer datasets, machine learning (ML) is widely recognized as the methodology of choice in Breast Cancer pattern classification and forecast modelling. All this is made very easy to implement in the Anaconda and Django framework which is used in the project.

There are datasets that are provided by UCI data repository (The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms) for Machine Learning for Breast Cancer diagnosis.

The datasets are real time datasets generated from patient records and they have around 76 attributes including a prediction attribute, which suggests regarding the severity of disease on scale of 1 to 4 or no presence at all. This database contains 76 attributes but for our experiment purposes we are using best 30 of those attributes. All attributes are numeric, and comma separated values. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

II. Proposed system

In the proposed work we have taken the one disease dataset from UCI machine learning repository for the classification. We have classified the attributes of disease dataset by applying classification algorithm. For evaluating the accuracy of all classification algorithms, the accuracy of the algorithm is determined for every attribute. Performance is determined on the basis FP rate, TP rate, recall, precision etc. Use training set data mode is used for training and testing the dataset.

In this proposed work following steps are performed.

There are many different algorithms as we are taking one input disease dataset:

- Input disease datasets (breast cancer)
- Select the attributes of disease datasets [for breast cancer 10 attributes]
- Pre-processing of all attributes by applying filter (add classification zeroR classifier) on the attributes.
- Visualize all the attributes of disease dataset.
- Select the data mining classification algorithms for creation of the classifier.
- Train and testing all the parameter for classification.
- Prediction and classification of disease is done.
- Evaluates results and analysis the performance of classifiers.

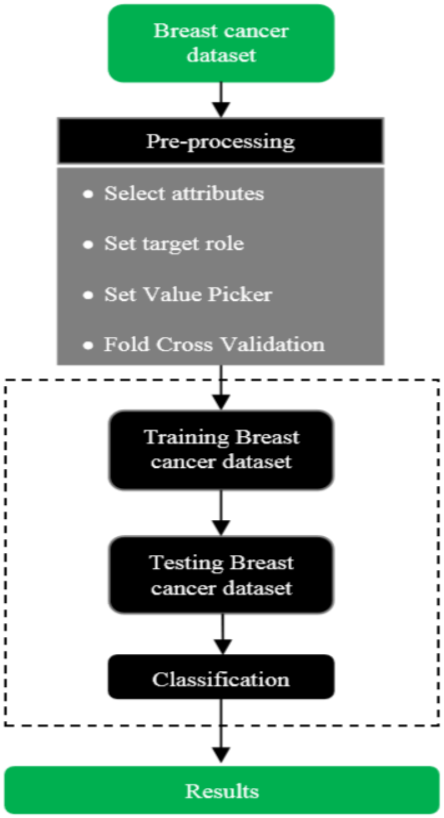


Fig1. Basic model architecture

As we have shown in the figure 1 that, at first we collected dataset of diagnosed cancer patients and by feature extracting we removed the irrelevant data which is not useful like patient id or name etc. after preprocessing of data we have trained 80% of data and remaining 20% is tested for classification using 5 different algorithms and compared the performance of each based on the correctness and specificity.

Based on the accuracy we got of different algorithms like Support Vector Machine(SVM), Logistic Regression(LR),k-Nearest neighbors(k-NN), Random Forest(RF) and Naïvye Bayes(NB) SVM is giving the highest accuracy of 97.43%. after checking the performance of every algorithm, the model predicts weather the cancer cell is benign or malignant.

III. Feature extraction part

The collected dataset of breast cancer patients is raw and unprocessed. That is dataset contains many attributes that is not useful for any kind, which in turn decreases the performance and disturbs the end results. So before feeding it to model we made sure that there are no null value attributes and the irrelevant like patient id etc. after performing feature extraction part the obtained dataset is new one with no unnecessary values is fed to the machine learning algorithms.

As we know most of the time our dataset contains features that vary lot in magnitude, units and range. As most of the ML algorithms use Euclidian distance between the two data points during computation. To bring all features to same magnitude scaling is used. That is transforming data so that fits into the scale of 0-100 or 0-1 like that.

IV. Selection of model

As it is the most important and exciting part To select a machine learning algorithm for new processed dataset is very much necessary to achieve better results in terms of accuracy and all. In our dataset our outcome variable having only two values, weather is the cancer cell is Benign(B) or Malignant(M).

Different classification algorithms used are:

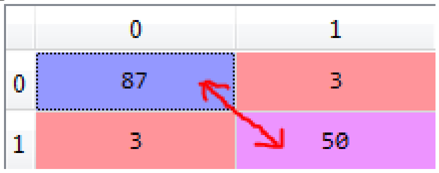
- LR—94.3%
- SVM—97.43%
- k-NN—95.1%
- RF—96.01%
- NB—92.41%

To find out we imported a confusion_matrix method of metric class. It is way of calculating number of misclassifications occurred based on true classification.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Fig2. Accuracy

To check the correct prediction, we have look at confusion matrix and add the predicted results diagonally which will be no. of correct prediction divided by total no. of predictions.



	0	1
0	87	3
1	3	50

Fig3.Confusion matrix

V. Conclusion and future work

The project is initiated because of a need of detection of cancer cells to treat and cure better efficient way. These requirements are given as a input file and then preprocessing is done. The next of modules managed to train the machine by 80% train data and the rest for testing purpose. Classifying ML algorithms are used to detect type of cells and result is stored.

Result is shown in both accuracy table and ROC curve for more clarity. It works with acceptable latency and their response times do not change with increase in number of input files. Thus, the system is efficient, reliable, consistent and user friendly.

For Future work:

To include more attributes as possible because the accuracy can be increased with increasing the attributes which are more useful for such cancer prediction.

VI. References

[1] L. Jena and N. K. Kamila, “Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease,” Int. J. Emerg. Res. Manag. &Technology, vol. 9359, no. 11, pp. 110–118, 2015.

[2] K. R. A. Padmanaban and G. Parthiban, “Applying Machine Learning Techniques for Predicting the Risk of Breast Cancer,” Indian J. Sci. Technol., vol. 9, no. 29, 2016.

[3] E. F. Hall, M., I. Wtten, Data mining: Practical machine learning tools and techniques, Kaufmann,. 2013.

[4] Arbab Masood Ahmad, Gul Muhammad, Khan, S.Ali Mahmud, Julian F. Miller, “Breast Cancer Detection Using Cartesian Genetic Programming evolved Artificial Neural Networks,” Philadelphia, Pennsylvania, USA, GECCO’12, July 7–11, 2016